COACTION PUBLISHING

Visualizing data quality: tools and views

Ian Painter[1]*, Julie Eaton[1], Don Olson[2], William Lober[1] and Debra Revere[1]

[1]University of Washington, Shoreline, WA, USA; [2]International Society for Disease Surveillance, Brighton, MA, USA

Objective

To present exploratory tools and methods developed as part of the data quality monitoring of Distribute data and discuss these tools and their applications with other participants.

Introduction

Distribute is a national emergency department syndromic surveillance project developed by the International Society for Disease Surveillance for influenza-like illness (ILI) that integrates data from existing state and local public health department surveillance systems. The Distribute project provides graphic comparisons of both ILI-related clinical visits across jurisdictions and a national picture of ILI.

Unlike other surveillance systems, Distribute is designed to work solely with summarized (aggregated) data, which cannot be traced back to the unaggregated 'raw' data. This and the distributed, voluntary nature of the project creates some unique data quality issues, with considerable site to site variability. Together with the ISDS, the University of Washington has developed processes and tools to address these challenges, mirroring work done by others in the Distribute community.

Methods

The University of Washington together with the ISDS has undertaken a comprehensive analysis of the quality of the data being received by Distribute, primarily using visual methods, examining data quality characteristics within and between sites. Several visualization tools were developed to assist in analyzing and characterizing data quality patterns for each site: upload pattern graphs (Fig. 1), stacked lag histograms and arrays of lagged time series graphs. Upload pattern graphs are heat maps comparing upload dates with encounter dates (an example figure is given below for three sites). Stacked lag histograms provide a succinct view of the complete distribution of data timeliness for a particular site. Arrays of lagged time series graphs provide an in-depth look at how timeliness patterns manifest in time series graphs. Implementation of the latter two visualizations required implementing a specific database architecture to enable reconstruction of the data at any prior upload date.

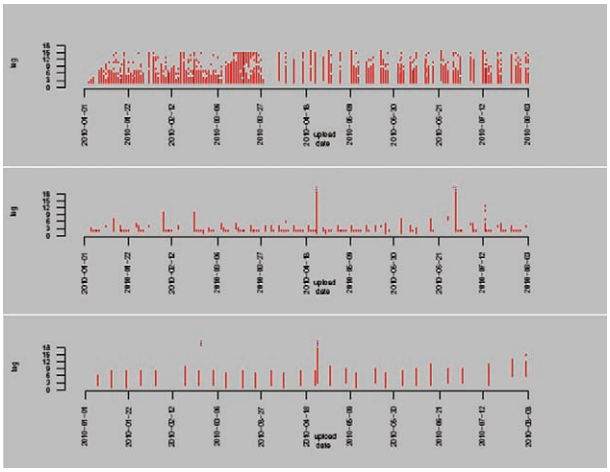

Fig. 1. Upload pattern plots for three sites. Each point represents an encounter date contained with in an upload file. The x-axis represents the date of the file upload and the y-axis represents the number of days prior to the upload date the encounter date was. The graph is truncated at 16 days prior to the upload date.

Results

In our talk, we will present these visualizations and demonstrate how they can be used to discover several common and some unusual data quality patterns and issues. We will also discuss the underlying architecture that allows us to reconstruct prior views and discuss the importance of examining data quality in terms of prior data views.

Keywords

Visualizations; data quality; surveillance

*Ian Painter
E-mail: ipainter@uw.edu

 Citation: Emerging Health Threats Journal 2011, **4**: 11144 - DOI: 10.3402/ehtj.v4i0.11144